Alternative Approaches for Cross-Language Text Retrieval

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

Introduction

The explosive growth of the Internet and other sources of networked information have made automatic mediation of access to networked information sources an increasingly important problem. Much of this information is expressed as electronic text, and it is becoming practical to automatically convert some printed documents and recorded speech to electronic text as well. Thus, automated systems capable of detecting useful documents are finding widespread application.

With even a small number of languages it can be inconvenient to issue the same query repeatedly in every language, so users who are able to read more than one language will likely prefer a multilingual text retrieval system over a collection of monolingual systems. And since reading ability in a language does not always imply fluent writing ability in that language, such users will likely find cross-language text retrieval particularly useful for languages in which they are less confident of their ability to express their information needs effectively.

The use of such systems can be also be beneficial if the user is able to read only a single language. For example, when only a small portion of the document collection will ever be examined by the user, performing retrieval before translation can be significantly more economical than performing translation before retrieval. So when the application is sufficiently important to justify the time and effort required for translation, those costs can be minimized if an effective cross-language text retrieval system is available. Even when translation is not available, there are circumstances in which cross-language text retrieval could be useful to a monolingual user. For example, a researcher might find a paper published in an unfamiliar language useful if that paper contains references to works by the same author that are in the researcher's native language.

Multilingual text retrieval can be defined as selection of useful documents from collections that may contain several languages (English, French, Chinese, etc.). This formulation allows for the possibility that individual documents might contain more than one language, a common occurrence in some applications. Both cross-language and within-language retrieval are included in this formulation, but it is the cross-language aspect of the problem which distinguishes multilingual text retrieval from its well studied monolingual counterpart. At the SIGIR 96 workshop on "Cross-Linguistic Information Retrieval" the participants discussed the proliferation of terminology being used to describe the field and settled on "Cross-Language" as the best single description of the salient aspect of the problem. "Multilingual" was felt to be too broad, since that term has also been used to describe systems able to perform within-language retrieval in more than one language but that lack any cross-language capability. "Cross-lingual" and "cross-linguistic" were felt to be equally good descriptions of the field, but "cross-language" was selected as the preferred term in the interest of standardization. Unfortunately, at about the same time the U.S. Defense Advanced Research Projects Agency (DARPA) introduced "translingual" as their preferred term, so we are still some distance from reaching consensus on this matter.

A couple of preliminary remarks are in order to establish the scope of this survey. My goal is to identify for you what I see as the key themes in cross-language text retrieval research in order to establish a common framework for our discussions during this symposium. I have not tried to describe each technical approach in detail because we will hear from many of the people whose work I cite over the course of the symposium. But I have endeavored to provide sufficient detail to illustrate the relationships between the different approaches that we will hear about at this symposium. Readers interested in a more detailed comparison of the work I was aware of a year ago may find my technical report on this subject useful (Oard & Dorr 1996b)[1]

Traditionally "text retrieval" and "information retrieval" have been used interchangeably, but as retrieval from other modalities (e.g., speech or images) has become more practical it is becoming more common to be explicit about the sort of information being retrieved. I have been careful to limit the scope of this survey to "text retrieval" because Peter Schauble and David James will provide the necessary background for our discussions about cross-language speech retrieval in their presentations.

I will not attempt to draw a sharp distinction between retrieval and filtering in this survey. Although my own work on adaptive cross-language text filtering has led me to make this distinction fairly carefully in other presentations (c.f., (Oard 1997b)), such an approach does little to help understand the fundamental techniques which have been applied or the results that have been obtained in this case. Since it is still common to view filtering (detection of useful documents in dynamic document streams) as a kind of retrieval, I will simply adopt that perspective here.

Controlled Vocabulary Retrieval

Cross-language text retrieval has an extensive research heritage. The first practical approach to cross-language text retrieval required that the documents be manually indexed using a predetermined vocabulary and that the user express the query using terms drawn from that same vocabulary. This is referred to as a "controlled vocabulary" approach. In such systems, a multilingual thesaurus is used to relate the selected terms from each language to a common set of language-independent concept identifiers, and document selection is based on concept identifier matching. In the hands of a skilled user who is familiar with controlled vocabulary search techniques, such systems can be remarkably effective. Of particular note, if well designed, controlled vocabulary cross-language text retrieval systems can be just as effective as monolingual applications of similar techniques.

Controlled vocabulary cross-language text retrieval systems are presently widely used in commercial and government applications for which the number of concepts (and hence the size of the indexing vocabulary) is manageable. Dagobert Soergel will present techniques for cross-language controlled vocabulary text retrieval in more detail (Soergel 1997), so I will limit my discussion of those issues here to identifying the limitations of controlled vocabulary techniques which have motivated the exploration of alternative approaches.

In many applications, the most challenging aspect of employing a controlled vocabulary technique is that terms from that vocabulary must be assigned to each document in the collection. This assignment of "descriptors" was traditionally done manually, but Marjorie Hlava will describe a semi-automated approach which can enable application of controlled vocabulary techniques to larger document collections than was previously practical (Hlava *et al.* 1997). High-volume applications such as filtering newswire stories and applications in which the documents are generated from diverse sources that are not easily standardized will, however, challenge the capabilities of even existing semi-automatic approaches. And low-cost applications such as cross-language World Wide Web search engines will demand fully automatic techniques which nonetheless must achieve reasonable performance.

A second important concern about controlled vocabulary text retrieval is that it has proven to be fairly difficult to train users to effectively select search terms and to exploit thesaurus relationships. In part this results from the limited capabilities of earlier user interfaces, an issue being investigated by Pollitt and his colleagues at the University of Huddersfield (Pollitt & Ellis 1993). Another approach to mitigating the effects of this problem is to embed some thesaurus navigation functionality into the retrieval engine itself. This is the approach being investigated by Richard Marcus at MIT (Marcus 1994). I am not as familiar with all of the work in this area as I would like to be, but it is my impression that we are a long way from developing retrieval interfaces that are as simple as those offered by modern web search engines. And so long as more complex interfaces (and hence, more sophisticated users) are needed to extract the full power from a system, an aggressive research program investigating alternative approaches to cross-language text retrieval seems well justified.

Free Text Retrieval

For text retrieval, the alternative to a controlled vocabulary is to use the words which appear in the documents themselves as the vocabulary. Such systems are referred to as "free text" (or sometimes "full text") retrieval systems. Two basic approaches to cross-language free text retrieval have been emerged: corpus-based approaches and knowledge-based approaches. These two approaches are not mutually exclusive, however, and the trend in cross-language free text retrieval research is to combine aspects of each to maximize retrieval effectiveness. Figure 1 illustrates the taxonomy

[1] Much of the source material cited in these surveys can can be obtained using my Cross-Language Text Retrieval web page at http://www.ee.umd.edu/medlab/mlir/

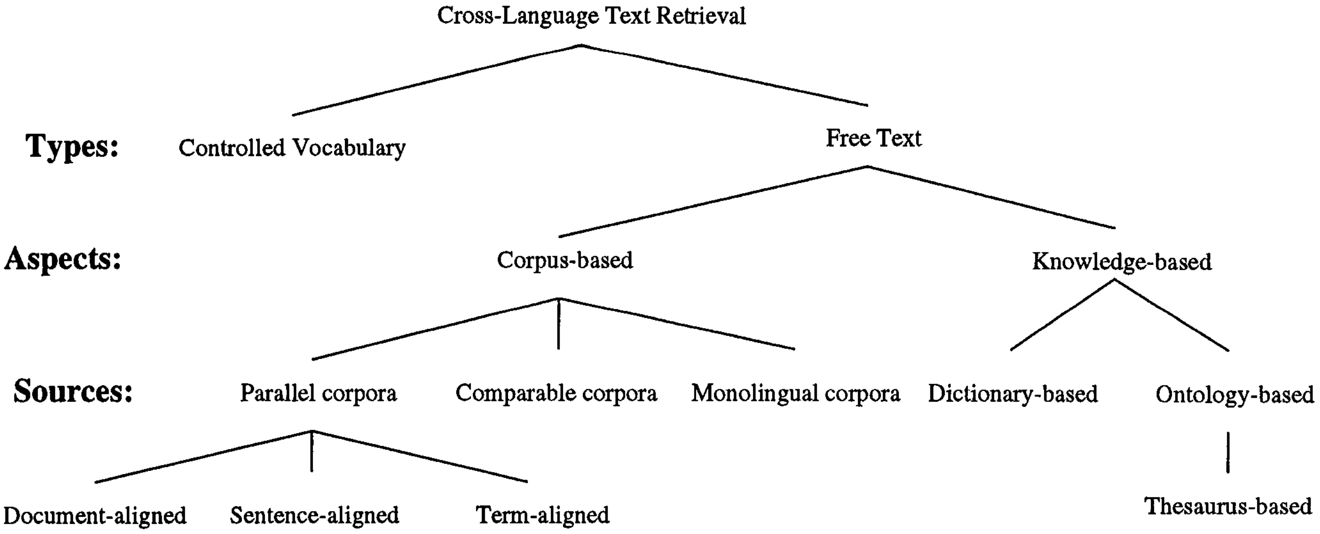

Figure 1: Cross-Language Text Retrieval Approaches

that I have described so far, and shows some useful distinctions between the sources of information that are exploited by different corpus-based and knowledge-based approaches.

In controlled vocabulary cross-language text retrieval it was essentially necessary to translate both the documents and the query into a common language, the controlled vocabulary itself. The assignment of descriptors is essentially a document translation process, and the process of forming queries using the controlled vocabulary is a (sometime entirely manual) query translation process. With free text retrieval, it becomes possible to consider approaches in which only the query or only the documents are translated. In most applications it would be more efficient to translate only the queries because queries are often rather short. This choice is so pervasive in the literature that I will implicitly assume a query translation strategy in the following discussion except when discussing specific techniques that are based on some other strategy.

Multilingual thesauri of the type used in controlled vocabulary text retrieval are one type of knowledge structure, and the knowledge-based approaches I will describe seek to apply multilingual thesauri and similar types of knowledge structures to free text retrieval. Bilingual dictionaries that were originally developed for human use are presently the most widely available cross-language knowledge structures that have the breadth of coverage required by many cross language free text retrieval applications, and several are available in electronic form. Thus, dictionary-based retrieval is the best explored branch of the knowledge-based cross-language text retrieval hierarchy.

The basic idea in dictionary-based cross-language text retrieval is to replace each term in the query with an appropriate term or set of terms in the desired language. Two factors limit the performance of this approach. The first is that many words do not have a unique translation, and sometimes the alternate translations have very different meanings. Monolingual text retrieval systems face similar challenges from homonomy and polysemy (multiple meanings for a single term), but translation ambiguity significantly exacerbates the problem. Use of every possible translation, for example, can greatly expand the set of possible meanings because some of those translations are likely to introduce additional homonomous or polysemous word senses in the second language. This problem is particularly severe in view of the observed tendency of untrained users to enter such short queries (often a single word) that it would not even be possible for a human to determine the intended meaning (and hence the proper query translation) from the available context.

The second challenge for a dictionary-based approach is that the dictionary may lack some terms that are essential for a correct interpretation of the query. This may occur either because the query deals with a technical topic which is outside the scope of the dictionary or because the user has entered some form of abbreviation or slang which is not included in the dictionary. As dictionaries specifically designed for query translation are developed, the effect of this limitation may be reduced. But it is unlikely to be eliminated completely because language use is a creative activity, with new terms entering the lexicon all the time. There will naturally be a lag between the introduction of a term and its incorporation into a standard reference work such as a dictionary. Of course, this last point applies equally well to controlled vocabulary systems based on multilingual thesauri since the introduction of a new term may have been motivated the the need to describe a new concept that did not previously appear in the document collection.

Of course, knowledge-based approaches may seek to exploit more sophisticated knowledge structures ("ontologies") as well. A thesaurus is one type of ontology, one specialized to information retrieval, and Dagobert Soergel will be discussing the use of multilingual thesauri by cross-language free text retrieval systems as well (Soergel 1997). Another approach with offers the potential for significant economies of scale is to exploit ontologies which are designed to support tasks beyond information retrieval as well. Jose Gilarranz will discuss the potential for using the EuroWordNet ontology in this way (Gilarranz, Gonzalo, & Verdejo 1997).

Corpus-based approaches seek to overcome the limitations of knowledge-based techniques by constructing query translation methods which are appropriate for the way language is used in a specific application. Because it can be impractical to construct tailored bilingual dictionaries or sophisticated multilingual thesauri manually for large applications, corpus-based approaches instead analyze large collections of existing text and automatically extract the information needed to construct automatic application-specific translation techniques. The collections which are analyzed may contain existing translations and the documents that were translated (a "parallel" collection), or they may be composed of documents on similar subjects which are written in different languages (a "comparable" collection).

Techniques Which Exploit Parallel Corpora

In 1990, Landauer and Littman (then with Bellcore) developed a corpus-based cross-language free text retrieval technique which has come to be known as Cross-Language Latent Semantic Indexing (CL-LSI) (Landauer & Littman 1990; 1991). Susan Dumais will present the latest results on CL-LSI at this symposium (Dumais *et al.* 1997). The remarkable thing about this work is that in addition to beginning the present development of cross-language free text retrieval, it remains to this day the only technique that has demonstrated cross-language free text retrieval effectiveness that is on a par with the within-language performance of that same technique (Dumais, Landauer, & Littman 1996). This result is particularly significant because a monolingual text retrieval system based on Latent Semantic Indexing has achieved effectiveness measures nearly equal to those of the best participating systems at the third Text Retrieval Conference (Dumais 1994).

In CL-LSI a set of representative bilingual documents are first used to form a training collection by adjoining a translation of each document to the document itself. A rank revealing matrix decomposition (the singular value decomposition) is then used to compute a mapping from sparse term-based vectors (usually with weights base on both within-document and collection-wide term frequency) to short but dense vectors that appear to capture the conceptual content of each document while suppressing the effect of variations in term usage. CL-LSI appears to achieve it's effectiveness by suppressing cross-language variations in term choice as well. In principle this technique can be extended to multiple languages, although the retrieval effectiveness of such a configuration has not yet been determined experimentally. Berry and Young repeated this work using passages from the Bible in English and Greek (Berry & Young 1995). They were able to demonstrate that fine-grained training data, using only the first verse of each passage to identify the principal components, improved retrieval performance over Landauer and Littman's coarser approach.

It is important to caveat the reported results for LSI by observing that both sets of experiments were conducted with experiment designs that matched the retrieval application to the characteristics of the parallel document collection that was used to develop the translation technique. Our experiments with this technique show a significant reduction in performance when a parallel document collection that is more weakly related to the retrieval application is used (Oard 1997a). This limitation is not unique to CL-LSI, however. It results from the fact that corpus-based techniques generally seek to balance the adverse effect of invalid inferences that result from misleading statistical cooccurrence observations with the beneficial effects of correctly recognizing that only a limited number of senses for words with several possible meanings are present in the training collection. As term use in the training and evaluation collections begins to diverge, this "beneficial" effect rapidly becomes a liability.

Mark Davis of New Mexico State University has conducted some large-scale cross-language text retrieval evaluations which shed some light on this limitation using material from the Text REtrieval Conferences (TREC) and he will be reporting on those results here (Davis 1996; Davis & Ogden 1997). Davis' recent TREC-5 experiments used 25 queries translated manually from Spanish to English and 173,000 Spanish language newswire stories and achieved about 75% of the average precision established in a monolingual evaluation using the same system and collection.

Davis' TREC-5 results indicate that when used alone, dictionary-based query expansion achieves about 50% of the average precision that would be achieved by a monolingual system, but that when translation ambiguity is limited this performance can be improved. This is quite consistent with similar results that have been obtained on smaller collections (Hull & Grefenstette 1996), suggesting that although the size of the collection may affect absolute performance measures, the effect on relative performance between monolingual and cross-language retrieval may be less significant. To improve over this baseline, Davis limited dictionary-based query expansion using part-of-speech information that was determined statistically, and combined this with additional constraints on the permissible translations that were determined using a large parallel corpus. This work is particularly interesting because it effectively combines dictionary-based and corpus-based techniques in a single retrieval system. And because the content of the parallel corpus of United Nations documents that was used was not particularly closely related to the content of the newswire stories, these experiments offer some insight into the effect of a domain mismatch as well.

Davis' corpus-based approach for restricting translation ambiguity seeks to select translations which would select similar sentences from documents in the parallel document collection. The technique is based on similarity matching between a vector which represents the query and vectors which represent individual sentences in the document collection. Thus, Davis is exploring a technique based on sentence-level alignment in a parallel collection in contrast to the coarser document-level alignment on which CL-LSI is based.

At the University of Maryland, Dorr and I have developed a technique based on term-level alignment which also offers the potential for integration of dictionary-based and corpus-based techniques (Oard 1996). The basic idea is to estimate the domain-specific probability distribution on the possible translations of each term based on the observed frequency with which terms align in a parallel document collection. We then use this statistically enhanced bilingual dictionary as a linear operator to rapidly map the vectors which represent documents from one language into another. The effectiveness of this technique depends on a sort of "consensus translation effect" in which several terms in the source language can potentially contribute to the weight assigned to a single term in the target language. As a result, it is only practical to apply our vector translation technique to vectors which represent documents. Typical queries simply don't contain enough terms or enough variation in term usage to develop a useful consensus translation effect. This limitation fits well with our focus on cross-language text filtering because our adaptive information need representation is not amenable to query translation.

In our initial experiments we have used a purely corpus-based approach for developing our statistically enhanced dictionary. In an evaluation conducted using the same collections used by Davis (and one additional TREC collection), we found that that implementation of our technique achieves about half the effectiveness of CL-LSI. An examination of the transfer mapping developed from our term alignment step reveals that many of the detected alignments do not represent valid translations. This is not a surprising result since term alignment is a challenging problem which is presently the focus of a good deal of research effort. In our next experiments we plan to constrain the allowable translations to those which occur in a broad-coverage bilingual dictionary, seeking to match or exceed the performance of CL-LSI.

In addition to demonstrating three techniques for adaptive cross-language text filtering, our work at Maryland has made two other contributions that may be of interest. The first is that we have developed a methodology for evaluating corpus-based adaptive cross-language text filtering effectiveness which does not depend on the development of an expensive specialized test collection (Oard & Dorr 1996a). A fair evaluation of such techniques requires two training collections, one of which must be a parallel bilingual corpus, and one evaluation collection. We have found a way to align TREC topic descriptions that were originally developed independently for each language and then to measure the quality of that alignment. Our approach is based on a very small number of topics, but until a suitable test collection is available for cross-language filtering evaluation it represents the best technique I know of for conducting such evaluations. A second useful result is that we have developed a technique to measure the degradation in effectiveness which results from the different domains of the UN collection and the Spanish documents used in TREC. This may be helpful when interpreting Davis' results, and more broadly it may offer some insight into the fundamental limits on the performance of corpus-based techniques when a well-suited parallel document collection is not available.

Alternatives to Parallel Corpora

Corpus-based approaches which exploit parallel document collections are limited by the requirement to obtain a suitable document collection before the technique can be applied. Of course, this is true of any corpus-based approach, but it poses a particularly severe problem for techniques based on parallel corpora. The translated documents are expensive to create, so the required translations are only likely to be available in highly specialized application domains. While a corpus-based technique developed from a parallel document collection can in principle be used for unrelated applications as well, significant reductions in retrieval effectiveness should be expected and there are experimental results which suggest that such performance penalties actually occur.

Techniques based on comparable document collections may eventually overcome this limitation, and Peter Schauble has described one such method at this symposium at this symposium (Sheridan, Wechsler, & Schäuble 1997). A comparable document collection is one in which document are aligned based on the similarity between the topics which they address rather than because they are direct translations of each other. The raw material for a comparable document collections is far easier to obtain that is the translated text required used in a parallel collection, but alignment of the individual documents is still a challenging task. Existing automatic and semiautomatic document alignment techniques are fairly application-specific, and considerable work likely remains to be done before it will be clear whether easily generalized techniques can be developed.

Perhaps the most intriguing alternative in this regard is the approach that Lisa Ballesteros of the University of Massachusetts will describe which does not require that the corpora be aligned at all (Ballesteros & Croft 1996; 1997). By exploiting a pseudo-relevance feedback technique that has been shown to be effective for within language retrieval, significant performance improvements over unconstrained dictionary-based query translation were achieved. This approach essentially seeks to modify the query to more closely resemble the documents in the collection. The University of Massachusetts team achieved their best results when performing this technique twice, once before the dictionary-based query translation and once before using the translated query to rank order the documents in the evaluation collection. Their technique requires the availability of document collections in each language, but it is not necessary that the individual documents in these collections be related in any way.

Lessons from Corpus Linguistics

The field of "corpus linguistics" has explored the use of corpus-based techniques in to a variety of applications such as text retrieval, speech recognition, machine translation and ontology construction. In each field the initial corpus-based experiments typically emphasize statistical analysis over linguistic theory, an approach which has led to some remarkable successes. In machine translation, for example, early statistical approaches demonstrated performance that was competitive with that achieved by contemporaneous linguistically motivated approaches (Brown *et al.* 1990). But purely statistical approaches also introduce errors that no human would make because the techniques typically exploit term cooccurrence and a fairly complex set of factors actually interact to produce these cooccurrences. The present statistical models are inadequate to capture some of these interactions, but significant performance improvements can be achieved when appropriate linguistically motivated constraints are effectively integrated with the statistical analysis. The experience to date with cross-language text retrieval suggests that similar improvements can be obtained for this aspect of corpus linguistics as well.

There is now mounting evidence from both corpus linguistics and cross-language text retrieval research that treating both individual words and multi-word phrases as the "terms" which are manipulated can significantly improve the effectiveness of cross-language techniques. van der Eijk observed this effect with adjacency-based phrases in automatic thesaurus construction experiments, Hull and Grefenstette observed it with dictionary-based phrases in cross-language text retrieval experiments, and Radwan and Fluhr observed it with dictionary-based phrases in cross-language text retrieval experiments that integrated both dictionary-based and corpus-based techniques (Hull & Grefenstette 1996; Radwan & Fluhr 1995; van der Eijk 1993). This is a particularly surprising result since the preponderance of the evidence on text retrieval in a single language indicates that multi-word phrases are of little use. Presumably the basis for this effect is that it is translation ambiguity which causes cross-language full-text retrieval systems to achieve lower retrieval effectiveness than their monolingual counterparts, and the use of phrases constrains this translation ambiguity to a significant extent. We are fortunate to have representatives of two of these three research groups with us. Hopefully this symposium will offer us the opportunity to explore this issue in some detail.

Corpus linguistics has also produced several useful tools for designers of cross-language text retrieval systems, many of which will be described at this symposium. One in particular that will be needed in almost every cross-language text retrieval application is language identification. With the notable exception of CL-LSI, cross-language text retrieval techniques typically require that the language in which the query and each document are expressed be known so that the correct processing can be applied. Fortunately, language identification techniques with better than 95% accuracy are now available (Kikui 1996).

System Integration

Although the techniques I have described span the range of technical approaches to the cross-language text retrieval problem, the application of these techniques to solve practical problems requires a considerably broader perspective. Yoshihiko Hayashi will describe an excellent example of end-to-end integration in cross-language text retrieval application at this symposium (Hayashi, Kikui, & Susaki 1997). The system, known as TITAN, allows users to enter queries to a web search engine in either English or Japanese. Statistical language identification is used to determine the languages used in each indexed web page, and page titles are translated into the language of the query using a translation technique that is tuned for the particular characteristics of web page titles. This last feature is particularly interesting, because it begins to address the potential of human-in-the-loop cross-language text retrieval processing. Robert Frederking will explore these issues in more detail in the next session (Frederking *et al.* 1997), and Megumi Kameyama will describe some advanced techniques for presenting useful information to the user later in the symposium (Kameyama 1997).

End-to-end solutions need not be complex, however, in order to be useful. An example of a simple but elegant implementation which exploits presently available technology is Paracel's Fast Data Finder, a text filtering system based on special purpose parallel processing hardware.[2] Provisions are provided in the Fast Data Finder to translate explicit information need specifications (profiles) between a limited number of languages using a dictionary-based technique. Users are then allowed to fine-tune the profile in each language. If sufficient domain and language expertise is available to allow refined profiles to be developed over time, such an approach offers the potential for fast and effective filtering of documents in multiple languages. Although the profile translation capability produces some cross-language retrieval functionality, it is the additional step (manual tuning) that makes it possible in this case to achieve optimum performance.

Conclusions

My goal in this presentation has been to provide a common framework for our discussion of cross-language text retrieval. The taxonomy I have described is only one way of depicting what is actually a fairly complex faceted classification scheme, but it provides a fairly clear way to describe where each approach that we will discuss fits into the range of possibilities that have been explored. I am hopeful that our colleagues interested in speech retrieval will also find this taxonomy useful for their purposes when considering the potential for cross-language speech retrieval systems.

This is an exciting time to be working on cross-language text retrieval. This symposium caps an eight month period that has seen five workshops around the world which have addressed various aspects of cross-language text retrieval, and this year for the first time TREC will include a cross-language text retrieval evaluation. I am particularly impressed that it appears we will have representatives from a majority of the research groups throughout the world that are actively working on this problem when we meet at Stanford. This forum thus offers us an unprecedented opportunity to forge the kind of international partnerships that I suspect will be needed to meet the demands of the worldwide market for cross-language text retrieval systems. I look forward to working with you to discern the most important issues that demand our attention and to identify promising directions for future research on this important topic.

Acknowledgments

The author would like to thank Bonnie Dorr, Dagobert Soergel and Philip Resnik for the extensive discussions which helped me to see these issues more clearly.

References

Ballesteros, L., and Croft, B. 1996. Ductionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, 791–801. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html.

Ballesteros, L., and Croft, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu//medlab/filter/sss/ /papers/ballesteros.ps.

[2] Paracel Inc., 80 South Lake Avenue, Suite 650, Pasadena, CA 91101-2616

Berry, M., and Young, P. 1995. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities* 29(6):413–429.

Brown, P. F.; Cocke, J.; Pietra, S. A. D.; Pietra, V. J. D.; Jelinek, F.; Lafferty, J. D.; Mercer, R. L.; and Roossin, P. S. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2):79–85.

Davis, M. W., and Ogden, W. C. 1997. Implementing cross-language text retrieval systems for large-scale text collections and the world wide web. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/davis.ps.

Davis, M. 1996. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In Harman, D. K., ed., *The Fifth Text REtrieval Conference (TREC-5).* NIST. To appear.

Dumais, S. T.; Letsche, T. A.; Littman, M. L.; and Landauer, T. K. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/dumais.ps.

Dumais, S. T.; Landauer, T. K.; and Littman, M. L. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In Grefenstette, G., ed., *Working Notes of the Workshop on Cross-Linguistic Information Retrieval.* ACM SIGIR. http://superbook.bellcore.com/~std/papers/SIGIR96.ps.

Dumais, S. T. 1994. Latent Semantic Indexing (LSI): TREC-3 report. In Harman, D., ed., *Overview of the Third Text REtrieval Conference*, 219–230. NIST. http://potomac.ncsl.nist.gov/TREC/.

Frederking, R.; Mitamura, T.; Nyberg, E.; and Carbonell, J. 1997. Translingual information access. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/frederking.ps.

Gilarranz, J.; Gonzalo, J.; and Verdejo, F. 1997. An approach to conceptual text retrieval using the eurowordnet multilingual semantic database. In *AAAI Symposium on Cross-Language Text and Speech retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/gilarranz.ps.

Hayashi, Y.; Kikui, G.; and Susaki, S. 1997. Titan: A cross-linguistic search engine for the www. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/haysahi.ps.

Hlava, M. M. K.; Hainebach, R.; Belonogov, G.; and Kuznetsov, B. 1997. Cross-language retrieval - English/Russian/French. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.edu/medlab/filter/sss/papers/hlava.ps.

Hull, D. A., and Grefenstette, G. 1996. Experiments in multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* http://www.xerox.fr/grenoble/mltt/people/hull/papers/sigir96.ps.

Kameyama, M. 1997. Information extraction across linguistic barriers. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/kameyama.ps.

Kikui, G. 1996. Identifying the coding system and language of on-line documents on the internet. In *Sixteenth International Conference of Computational Linguistics (COLING).* International Committee on Computational Linguistics. http://isserv.tas.ntt.jp/chisho/paper/9608KikuiCOLING.ps.Z.

Landauer, T. K., and Littman, M. L. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research.* Waterloo Ontario: UW Centre for the New OED and Text Research. 31–38. http://www.cs.duke.edu/~mlittman/docs/x-lang.ps.

Landauer, T. K., and Littman, M. L. 1991. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Eleventh International Conference: Expert Systems and Their Applications*, volume 8, 77–85.

Marcus, R. S. 1994. Intelligent assistance for document retrieval based on contextual, structural, interactive Boolean models. In *RIAO 94 Conference Proceedings, Intelligent Multimedia Information Retrieval Systems and Management*, volume 2, 27–

43. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (C.I.D.).

Oard, D. W., and Dorr, B. J. 1996a. Evaluating cross-language text filtering effectiveness. In Grefenstette, G., ed., *Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop.* ACM SIGIR. http://www.ee.umd.edu/medlab/filter/papers/sigir96.ps.

Oard, D. W., and Dorr, B. J. 1996b. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

Oard, D. W. 1996. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications.* Ph.D. Dissertation, University of Maryland, College Park. http://www.ee.umd.edu/medlab/filter/papers/thesis.ps.gz.

Oard, D. W. 1997a. Adaptive filtering of multilingual document streams. In *Submitted to RIAO 97.*

Oard, D. W. 1997b. The state of the art in text filtering. *User Modeling and User Adapted Interaction.* To appear.

Pollitt, A. S., and Ellis, G. 1993. Multilingual access to document databases. In *21st Annual Conference Canadian Society for Information Science*, 128–140.

Radwan, K., and Fluhr, C. 1995. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, 121–136.

Sheridan, P.; Wechsler, M.; and Schäuble, P. 1997. Cross-language speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/sheridan.ps.

Soergel, D. 1997. Multilingual thesauri in cross-language text and speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Rerieval.* American Association for Artificial Intelligence. To appear. http://www.ee.umd.edu/medlab/filter/sss/papers/soergel.ps.

van der Eijk, P. 1993. Automating the acquisition of bilingual terminology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, 113–119.